Journées Ouvertes de Biologie Informatique et de Mathématiques (JOBIM)
Marseille, France, Juillet 3-6, 2018
Abstracts des Journées, Brun C and Ballester B (eds.), A601, P202, p. 618-621.

IMGT/HighV-QUEST statistical analysis of IMGT clonotypes (AA), novel interface and functionalities for NGS analysis of IG and TR

Marianne Lèbre *† [1], Karthik Kalyan‡ [1], Patrice Duroux§ [1], Véronique Giudicelli¶ [1], Sofia Kossida‖ [1], Marie-Paule Lefranc** [1]

[1] IMGT®, the international ImMunoGeneTics information system®, Laboratoire d'ImmunoGénétique Moléculaire (LIGM) – Institut de Génétique Humaine (IGH) CNRS Université de Montpellier UMR 9002 – 141 rue de la Cardonille 34396 Montpellier Cedex 5, France

Introduction

IMGT®, the international ImMunoGeneTics information system®, http://imgt.org/ [1], is the global reference in immunogenetics and immunoinformatics [2], founded in 1989 by Marie-Paule Lefranc at Montpellier (Université de Montpellier and CNRS). IMGT® is a high-quality integrated knowledge resource specialized in the immunoglobulins (IG) or antibodies, T cell receptors (TR), major histocompatibility (MH) of humans and other vertebrate species, and in the immunoglobulin superfamily (IgSF), MH superfamily (MhSF) and related proteins of the immune system (RPI) of vertebrates and invertebrates.

IG and TR are the antigen receptors of the adaptive immune response which characterizes the vertebrates with jaws (*gnasthostomata*) [2]. Their study in normal and pathological conditions is a challenge due to their huge diversity (1012 potential specificities for humans) only limited by the number of the B and T cells that an organism is genetically programmed to produce. The high diversity of the variable domain at the N-terminal end of each IG or TR chain results from genomic DNA rearrangements which occur in B or T cells, respectively, and which involve variable (V), diversity (D) and joining (J) genes [2]. This combinatorial V-(D)-J diversity is further increased by the junctional diversity and for IG, by somatic hypermutations. The analysis of the immune repertoires has become feasible thanks to the developments of the next generation sequencing (NGS) technologies. Since 2010, IMGT® has developed IMGT/HighV-QUEST [3, 4], the high-throughput version of IMGT/V-QUEST, which is so far the only online tool available on the Web for the direct analysis of complete IG and TR variable domains (V-DOMAIN, corresponding to the V-(D)-J REGION) of NGS nucleotide rearranged sequences, from humans and other vertebrate species.

[*] Speaker
[†] Corresponding author: marianne.lebre@igh.cnrs.fr
[‡] Corresponding author: karthik.kalyan@igh.cnrs.fr
[§] Corresponding author: Patrice.Duroux@igh.cnrs.fr
[¶] Corresponding author: veronique.giudicelli@igh.cnrs.fr
[‖] Corresponding author: sofia.kossida@igh.cnrs.fr
[**] Corresponding author: Marie-Paule.Lefranc@igh.cnrs.fr

METHODOLOGY

IMGT/HighV-QUEST analyzes up to 500,000 sequences per run, with the same degree of resolution and high-quality results as IMGT/V-QUEST and IMGT/JunctionAnalysis [3, 4]. Indeed IMGT/HighV-QUEST uses the same algorithm and runs against the same IMGT reference directories. IMGT/HighV-QUEST numbers the user sequences according to the IMGT unique numbering and introduces gaps accordingly. It identifies the V, D and J genes in rearranged IG and TR sequences and, for the IG, characterizes the nt mutations and amino acid (AA) changes resulting from somatic hypermutations by comparison with the IMGT/V-QUEST reference directories. The tool integrates IMGT/JunctionAnalysis for the detailed characterization of the V-D-J or V-J junctions, IMGT/Automat for a complete sequence annotation with the delimitation of the IMGT labels of description. By default, IMGT/HighV-QUEST identifies the insertions/deletions (indels) which are NGS errors resulting from homopolymer hybridization and corrects them. IMGT/HighV-QUEST results consists classically of 11 CSV text files downloadable as an archive file [3, 4]. The CSV files contain one line per analysed sequence, and together may comprise up to 539 columns for a complete results report.

The IMGT/HighV-QUEST statistical analysis, which allows the identification and characterization of the clonotypes [5], may analyse up to one million IMGT/HighV-QUEST results.

RESULTS

In the literature, clonotypes characterize the repertoires of the adaptive immune responses but are defined differently, depending on the experiment design (functional specificity) or available data. Thus, a clonotype may denote either a complete receptor (e.g., TR-alpha_beta), or only one of the two chains of the receptor (e.g., TRA or TRB), or one domain (e.g., V-BETA), or the CDR3 sequence of a domain. Moreover the sequence can be at the AA or nt level, and this is rarely specified. Therefore, IMGT priority was to define clonotypes and their properties, which could be identified and characterized by IMGT/HighV-QUEST within the statistical analysis, unambiguously. In IMGT, the clonotype, designated as 'IMGT clonotype (AA)', is defined by a unique V-(D)-J rearrangement (with IMGT gene and allele names determined by IMGT/HighV-QUEST at the nt level) and a unique CDR3-IMGT AA junction sequence [5]. An IMGT clonotype (nt) is defined by a unique V-(D)-J rearrangement (with IMGT gene and allele names determined by IMGT/HighV-QUEST at the nt level) and a unique CDR3-IMGT nt junction sequence. Several IMGT clonotypes nt may correspond to one IMGT clonotype (AA).

The statistical analysis applies a filter on the IMGT/HighV-QUEST results: only the ones characterized by a V-GENE and allele (single or several alleles), a JUNCTION and a J-GENE and allele (single or several alleles) are filtered-in for statistical analysis [5]. Statistical analysis output is provided as a txz file (IMGT/HighV-QUEST Documentation: http://www.imgt.org/HighV-QUEST/doc.action). In order to evaluate and to explore, between sets, the significance of pairwise comparison of IMGT clonotype (AA) diversity and expression per V, D and J gene, IMGT/StatClonotype [6] was developed. This tool is downloadable on the IMGT® site (http://www. Integrated in the R package "IMGTStatClonotype", it offers a graphical interface to visualize pair wise comparison, per IMGT genes and alleles, of the IMGT clonotype (AA) diversity or expression of any IG or TR immunoprofiles of any species, obtained as outputs of the IMGT/HighV-QUEST statistical analysis.

With the advent of single molecule, long read sequencing (PacBio), the advanced functionality "Analysis of single chain Fragment Variable (scFv)" sequences [7] has been added as an option within IMGT/HighV-QUEST. So far, the NGS analysis of scFv was a challenge. Indeed, scFv are engineered antibody single chain fragments which comprise two variable domains associated by a peptide linker and the NGS methods did not provide reads long enough to span the length of the scFv (> 800 bp). If selected, the IMGT/HighV-QUEST functionality analyses both V-DOMAIN individually (results in the 11 CSV files) and produces a 12th CSV result file "scFv" where the association between the two V-DOMAIN, their respective positions in the sequence, the positions and the length of the linker are recorded. This advanced functionality allows the analysis of the content of scFv combinatorial phage display libraries which are classically screened for identification of novel therapeutic antibody specificities.

CONCLUSIONS AND PERSPECTIVES

IMGT/HighV-QUEST is the standard for the NGS analysis of IG and TR repertoires in experimental engineered (combinatorial libraries) or in physiological conditions (vaccination, immunodeficiency, autoimmune diseases, cancers and infectious diseases). IMGT/HighV-QUEST is particularly well adapted for the analysis of complete V domains of the IG and TR repertoires from B and T subsets, in many experiments and from many individuals (humans or other vertebrate species).

IMGT/HighV-QUEST was originally developed using Java based technologies and was initially a 2-tier architecture system: application, database. It combined a web-based user interface (client UI) and a job management system (using Java Quartz) in one web application. It evolved into a 3-tier architecture system: client UI, database and scheduling-system. The scheduling-system is now a standalone system (shell scripts and cron) which has the possibility to be integrated to an automation tool, such as Rundeck. After the implementation of the 3-tier architecture system, a new client UI will soon be made available based on modern web technologies (Bootstrap, Struts2 and Tiles3). The 3-tier architecture will enable easier implementation of the newly developed functionalities.

IMGT/HighV-QUEST makes use of High Performance Computing (HPC) clusters to run large number of user submitted jobs that are split into tasks (i.e. IMGT/V-QUEST runs) based on a linear equation. In regards to the new available HPC clusters and the users' demands to increase the number of sequences in one job, the reduction in computational processing time by parallelizing the IMGT/V-QUEST module within IMGT/HighV-QUEST is underway with the usage of dynamic parameters. Java 'multi-threading (fork-join/work-stealing-algorithm) and profiling benchmarking (JMH APIs)' are used for its development. The possibility of utilizing distributed computing facilities (JPPF) is also explored.

ACCESS TO HPC RESOURCES

This work was granted access to the HPC@LR and to the HPC resources of the Centre Informatique National de l'Enseignement Supérieur (CINES) and to Très Grand Centre de Calcul (TGCC) of the Commissariat à l'Energie Atomique et aux Energies Alternatives (CEA) under the allocation [036029] (2010-2018) made by GENCI (Grand Equipement National de Calcul Intensif).

REFERENCES

Lefranc M-P, Giudicelli V, Duroux P, Jabado-Michaloud J, Folch G, Aouinti S, Carillon E, Duvergey H, Houles A, Paysan-Lafosse T, Hadi-Saljoqi S, Sasorith S, Lefranc G, Kossida S. IMGT®, the international ImMunoGeneTics information system® 25 years on. Nucl. Acids Res. 2015 Jan;43(Database issue):D413-22. doi: 10.1093/nar/gku1056. Epub 2014 Nov 5 Free article. PMID: 25378316

Lefranc M-P. Immunoglobulin (IG) and T cell receptor genes (TR): IMGT® and the birth and rise of immunoinformatics. Front Immunol. 2014 Feb 05;5:22. doi: 10.3389/fimmu.2014.00022. Open access. PMID: 24600447

Alamyar E., Giudicelli V. Duroux P, Lefranc, M.-P. *IMGT/HighV-QUEST: a high-throughput system and web portal for the analysis of rearranged nucleotide sequences of antigen receptors - High-throughput version de IMGT/V-QUEST*. Poster no27 (abstract no60). 11èmes Journées Ouvertes de Biologie, Informatique et Mathématiques (JOBIM 2010).Montpellier, France (7-9 septembre 2010).

http://www.sfbi.fr/sites/default/files/jobim/jobim2010/index59385938.html?q=fr/node/55#IMGTHighV-QUEST:_A_High-Throughput_System_and_Web_Portal_for_the_Analysis_of_Rearranged_Nucleotide_Sequences_
_High-Throughput_Version_of_IMGTV-QUEST_

Alamyar E, Giudicelli V, Shuo L, Duroux P, Lefranc M-P. IMGT/HighV-QUEST: the IMGT® web portal for immunoglobulin (IG) or antibody and T cell receptor (TR) analysis from NGS high throughput and deep sequencing. Immunome Res. 2012, April 20;8:1:2. doi: 10.4172/1745-7580.1000056. PMID: 22647994

Li S., Lefranc M.-P., Miles J.J., Alamyar E., Giudicelli V., Duroux P., Freeman J.D., Corbin V.D.A., Scheerlinck J.-P., Frohman M.A., Cameron P.U., Plebanski M., Loveland B., Burrows S.R., Papenfuss A.T., Gowans E.J. IMGT/HighV-QUEST paradigm for T cell receptor IMGT clonotype diversity and next generation repertoire immunoprofiling. Nat. Commun. 2013;4:2333. doi:10.1038/ncomms3333 Open access. PMID: 23995877

Aouinti S, Giudicelli V, Duroux P, Malouche D, Kossida S, Lefranc M-P. IMGT/StatClonotype for Pairwise Evaluation and Visualization of NGS IG and TR IMGT Clonotype (AA) Diversity or Expression from IMGT/HighV-QUEST. Front Immunol. 2016 Sep 9;7:339. doi: 10.3389/fimmu.2016.00339. eCollection 2016. Free PMC Article. PMID: 27667992

Giudicelli V, Duroux P, Kossida S, Lefranc M-P. IG and TR single chain Fragment variable (scFv) sequence analysis: a new advanced functionality of IMGT/V-QUEST and IMGT/HighV-QUEST. BMC Immunol. 2017 Jun 26;18(1):35. doi: 10.1186/s12865-017-0218-8. PMID: 28651553.

Keywords: IMGT, immunoinformatics, immunogenetics, immunoglobulin, antibody, receptor T, IMGT clonotype (AA), NGS